CS 134: Networks
Problem Set 7

Xiner Zhou

March 30, 2017

1. Estimating Marginal Contribution to Influence using Sampling. (30 points)

a. Solution: X_S is the random variable denoting the number of nodes infected when S is chosen as the initial set of adopters, i.e., $r_i(S) = |R_i(S)|$, with probability distribution of $P[G = G_i]$, where $G_i = (V, E_i)_{i=1}^m$ are realized graphs of the random graph $G = (V, E, p)$. Then, by definition, the expected number of nodes infected by the initial set of adopters is just the expectation of X_S:

$$f(S) = \sum_{i=1}^{m} P[G = G_i]|R_i(S)|$$

$$= \sum_{i=1}^{m} P[G = G_i]r_i(S)$$

$$= E[X_S]$$

b. Solution:

$$f_S(a) = f(S \cup a) - f(S) = E[X_{S\cup a}] - E[X_S] = E[X_{S\cup a} - X_S]$$

c. Solution: The empirical estimation of $f_S(a)$ is the average value of l samples $X_{i,S\cup a} - X_{i,S}{}_{i=1}^{l}$. Here, we want to bound the estimation error, so that, once we have enough samples, we will be guaranteed to have an estimation not too far from the truth.

We want $P[|\widetilde{f_S(a)} - f_S(a)| \leq \epsilon] \geq 1 - \frac{1}{n^2 k}$, it is equivalent to :

$$P[|\widetilde{f_S(a)} - f_S(a)| \geq \epsilon] \leq \frac{1}{n^2 k}$$

.

Since $\widetilde{f_S(a)}$ is just the empirical mean of l samples of $X_{S\cup a} - X_S$, using the Chernoff bound, we know that:

$$P[|\widetilde{f_S(a)} - f_S(a)| \geq \epsilon] \leq 2e^{\frac{-2l\epsilon^2}{n^2}}$$

Therefore, if $2e^{\frac{-2l\epsilon^2}{n^2}} \leq \frac{1}{n^2 k}$, then we are guaranteed to have:

$$P[|\widetilde{f_S(a)} - f_S(a)| \geq \epsilon] \leq e^{\frac{-2l\epsilon^2}{n^2}} \leq \frac{1}{n^2 k}$$

Given a precision parameter $\epsilon > 0$, the number of nodes in the network n, and the cardinality constraint k. Solve the inequality for sample size l:

$$2e^{\frac{-2l\epsilon^2}{n^2}} \leq \frac{1}{n^2 k}$$

$$\Rightarrow l \geq \frac{n^2}{2\epsilon^2} log(2n^2 k)$$

Now we have proved that for our influence function, we can obtain arbitraty good approximations of the marginal contribution of a node using a modest number of samples.

d. Solution: Given graph $G = (V, E)$ with edge probabilities $p_{v,w_{(v,w)\in E}}$, a precision parameter $\epsilon > 0$, the number of nodes in the network n, and the cardinality constraint k. Let the sample size $m \geq \frac{n^2}{2\epsilon^2} log(2n^2 k)$. For any set $S \subseteq V$:

Algorithm:

1 For $i = 1$ to m do:

[2] Realize every edge in $(v, w) \in E$ with probabilities $p_{v,w}$ and set E' to be the set of realized edges.

[3] For every node $a \notin S$: Calculate marginal contribution of a to the set S in the ith realization, denotes as $\widetilde{f_{i,S}(a)}$.

4 End for

5 Calculate the empirical mean of the $m's$ sample for all nodes $a \notin S$, as the estimated marginal contribution using sampling, i.e.,

$$\forall a \notin S, \widetilde{f_S(a)} = \frac{1}{m} \sum_{i=1}^{m} f_{i,S}(a)$$

.

6 Return the element a that maximze the estimated marginal contribution using sampling, that is,

$$a = arg\ max_{a \notin S} \widetilde{f_S(a)}$$

The algorithm guarantees that $P(\widetilde{f_S(a)} \geq max_{b \in V} f_S(b) - \frac{\epsilon^2}{k}) \geq 1 - \frac{1}{nk}$.

Proof:

$$P(\widetilde{f_S(a)} \leq max_{b \in V} f_S(b) - \frac{\epsilon^2}{k}) = P(\cup_{b \in V} \widetilde{f_S(a)} \leq f_S(b) - \frac{\epsilon^2}{k})$$

$$\leq \sum_{b \in V} P(\widetilde{f_S(a)} \leq f_S(b) - \frac{\epsilon^2}{k}) \text{ (1)}$$

By the definition of $\widetilde{f_S(a)} = max_{b \notin S} \widetilde{f_S(b)}$, we know that,

$$\widetilde{f_S(a)} \geq \widetilde{f_S(b)}, \forall b \in V$$

$$\Rightarrow \widetilde{f_S(a)} \leq f_S(b) - \frac{\epsilon^2}{k} \subseteq \widetilde{f_S(b)} \leq f_S(b) - \frac{\epsilon^2}{k}, \forall b \in V$$

$$\Rightarrow (1) \leq \sum_{b \in V} P(\widetilde{f_S(b)} \leq f_S(b) - \frac{\epsilon^2}{k}) \ (2)$$

From 1(c) we know that, if the sample size $l \geq \frac{n^2}{2(\frac{\epsilon^2}{k})^2} log(2n^2k)$, then for any node b,

$$P(|\widetilde{f_S(b)} - f_S(b)| \geq \frac{\epsilon^2}{k}) \leq \frac{1}{n^2k}$$

Since $|\widetilde{f_S(b)} - f_S(b)| \geq \frac{\epsilon^2}{k} = \widetilde{f_S(b)} - f_S(b) \leq -\frac{\epsilon^2}{k} \cup \widetilde{f_S(b)} - f_S(b) \geq \frac{\epsilon^2}{k}$

$$\Rightarrow P(\widetilde{f_S(b)} \leq f_S(b) - \frac{\epsilon^2}{k}) \leq \frac{1}{n^2k}$$

$$\Rightarrow (2) \leq n\frac{1}{n^2k} = \frac{1}{nk}$$

$$\Rightarrow P(\widetilde{f_S(a)} \leq max_{b \in V} f_S(b) - \frac{\epsilon^2}{k}) \leq \frac{1}{nk}$$

$$\Rightarrow P(\widetilde{f_S(a)} \geq max_{b \in V} f_S(b) - \frac{\epsilon^2}{k}) \geq 1 - \frac{1}{nk}$$

2. The greedy algorithm with approximate marginals. (30 points)

a. Solution:
Since we assume that for every $o \notin S$, $f_S(o) \geq \frac{\epsilon}{k}$

$$\Rightarrow -\frac{\epsilon}{k} \geq -f_S(o^*) \ (1)$$

Since $o^* \in V$,

$$max_{b \in V} f_S(b) \geq f_S(o^*) \ (2)$$

Therefore, if we have an element $a \in V$ that respects $f_S(a) \geq max_{b \in V} f_S(b) - \frac{\epsilon^2}{k}$

$$(1) + (2) \Rightarrow$$

$$f_S(a) \geq max_{b \in V} f_S(b) - \epsilon f_S(o^*)$$
$$\geq f_S(o^*) - \epsilon f_S(o^*)$$
$$= (1 - \epsilon) f_S(o^*)$$

b. Solution: We first prove Lemma 1 and Lemma 2, then use Lemma 2 to prove that:

$$f(S) \geq (1 - \frac{1}{e^{1-\epsilon}})OPT$$

.

Lemma 1: At every step $i \in 1, ..., k$ we have that: $f(S_{i+1}) - f(S_i) \geq \frac{1-\epsilon}{k}(f(O) - f(S_i))$.

Proof: Let $O = o_1, ..., o_l$, and o_{max} be the element with the highest marginal contribution in O at stage $i + 1$. That is: $o_{max} = argmax_{o \in O} f_{S_i}(o)$. At stage $i + 1$ the algorithm selects element a_{i+1} and we are guaranteed that its marginal contribution is the highest.

In particular, its marginal contribution is also higher than the marginal contribution of the element in O that has the highest marginal contribution: $f_{S_i}(a_{i+1}) \geq f_{S_i}(o_{max})$. At 2(a) we have proved that if $f_S(a) \geq max_{b \in V} f_S(b) - \frac{\epsilon^2}{k}$, then $f_S(a) \geq (1-\epsilon) f_S(o^*)$. Therefore:

$$f_{S_i}(O) \leq \sum_{i=1}^{k} f_{S_i}(o_j) \ (subadditivity)$$

$$\leq k f_{S_i}(o^*)$$

$$\leq k \frac{1}{1-\epsilon} f_{S_i}(a_{i+1})$$

$$= k \frac{1}{1-\epsilon} (f(S_i \cup a_{i+1}) - f(S_i))$$

$$= k \frac{1}{1-\epsilon} (f(S_{i+1}) - f(S_i))$$

We therefore have:

$$\Rightarrow f(S_{i+1}) - f(S_i) \geq \frac{1-\epsilon}{k} f_{S_i}(O)$$

$$= \frac{1-\epsilon}{k} (f(S_i \cup O) - f(S_i))$$

$$\geq \frac{1-\epsilon}{k} (f(O) - f(S_i))$$

as required.

Lemma 2: At every step $i \in 1, ..., k$ we have that: $f(S_i) \geq (1 - (1 - \frac{1-\epsilon}{k})^i) f(O)$.

Proof: The proof is by induction on i.

- For $i = 1$, using Lemma 1:

$$f(S_1) \geq \frac{1-\epsilon}{k} f(O) = (1 - (1 - \frac{1-\epsilon}{k})) f(O)$$

- We now assume the claim holds for $i = l$ and we will prove that it holds for $i = l + 1$.

$$f(S_{l+1}) \geq \frac{1-\epsilon}{k} (f(O) - f(S_l)) + f(S_l)$$

$$= \frac{1-\epsilon}{k} f(O) + (1 - \frac{1-\epsilon}{k}) f(S_l)$$

$$\geq \frac{1-\epsilon}{k} f(O) + (1 - \frac{1-\epsilon}{k})(1 - (1 - \frac{1-\epsilon}{k})^l) f(O)$$

$$= \frac{1-\epsilon}{k} f(O) + ((1 - \frac{1-\epsilon}{k}) - (1 - \frac{1-\epsilon}{k})^{l+1}) f(O)$$

$$= \frac{1-\epsilon}{k} f(O) - \frac{1-\epsilon}{k} f(O) + (1 - (1 - \frac{1-\epsilon}{k})^{l+1}) f(O)$$

$$= (1 - (1 - \frac{1-\epsilon}{k})^{l+1}) f(O)$$

Proof the Main Results:

For $\forall k \geq 1$,

$$(1 - \frac{1-\epsilon}{k})^k = (1 - \frac{1}{\frac{k}{1-\epsilon}})^{\frac{k}{1-\epsilon}1-\epsilon} \leq \frac{1}{e^{1-\epsilon}}$$

Using Lemma 2, we know that:

$$\Rightarrow f(S) \geq (1 - (1 - \frac{1-\epsilon}{k})^k)OPT \geq (1 - \frac{1}{e^{1-\epsilon}})OPT$$

3. Putting it all together. (20 points)

Solution: **Modified Greedy Algorithm**

Given graph $G = (V, E)$ with edge probabilities $p_{v,w_{(v,w)\in E}}$, a precision parameter $\epsilon > 0$, the number of nodes in the network n, and the cardinality constraint k. Let the sample size $m \geq \frac{n^2}{2\epsilon^2} log(2n^2k)$. For any set $S \subseteq V$:

1 Set $S = \emptyset$

2 While $|S| \leq k$ do

[3] For $i = 1$ to m do:

[4] Realize every edge in $(v, w) \in E$ with probabilities $p_{v,w}$ and set E' to be the set of realized edges.

[5] For every node $a \notin S$: Calculate marginal contribution of a to the set S in the ith realization, denotes as $\widetilde{f_{i,S}(a)}$.

[6] End for

[7] Calculate the empirical mean of the $m's$ sample for all nodes $a \notin S$, as the estimated marginal contribution using sampling, i.e.,

$$\forall a \notin S, \widetilde{f_S(a)} = \frac{1}{m} \sum_{i=1}^{m} f_{i,S}(a)$$

.

[8] Return the element a that maximze the estimated marginal contribution using sampling, that is,

$$a = arg\ max_{a \notin S} \widetilde{f_S(a)}$$

[9] Let $S = S \cup a$

10 End while loop

11 Return S.

Proof: In 1(d) we have proved that, for any $S \subseteq V$, the algorithm returns a node $a \notin S$, such that $P(f_S(a) \geq max_{b \in V} f_S(b) - \frac{\epsilon^2}{k}) \geq 1 - \frac{1}{nk}$. In 2(a) and 2(b), we then proved that, if $f_S(a) \geq max_{b \in V} f_S(b) - \frac{\epsilon^2}{k}$, then $f(S) \geq (1 - \frac{1}{e^{1-\epsilon}})OPT$. It is equivalent to that,

$$P(f(S) \geq (1 - \frac{1}{e^{1-\epsilon}})OPT) \geq P(f_S(a) \geq max_{b \in V} f_S(b) - \frac{\epsilon^2}{k}) \geq 1 - \frac{1}{n}$$

.

Note that for small values of $\epsilon > 0$, we have:

$$1 - \frac{1}{e^{1-\epsilon}} = 1 - \frac{1}{e}e^{\epsilon} \approx 1 - \frac{1}{e}(1 + \epsilon) = 1 = \frac{1}{e} - \frac{\epsilon}{e} \geq 1 - \frac{1}{e} - \epsilon$$

Therefore:

$$P(f(S) \geq (1 - \frac{1}{e} - \epsilon)OPT) \geq P(f(S) \geq (1 - \frac{1}{e^{1-\epsilon}})OPT)$$
$$\geq 1 - \frac{1}{nk}$$
$$\geq 1 - \frac{1}{n}$$

(since $k \geq 1$)

So we have proved, the set S of k nodes returned by the modified greedy algorithm satisfies: $P(f(S) \geq (1 - \frac{1}{e} - \epsilon)OPT) \geq 1 - \frac{1}{n}$.

4. Programming: Maximizing Influence on Networks (20 points)

a. Solution: The probability of node 42 influencing node 75 is 0.00593041258054.

b. Solution: The average number of nodes in the randomly realized graph is 187.75.

c. Solution: $f(S) = 4.198$

d. Solution:

Due to computation time (maybe my PC is too old and slow), I wasn't able to select 5 nodes as initial adopters, it took me more than 18 hours and still couldn't finish running. However, my code should be fine, since the program doesn't report any error and kept printing out the intermediate results. Please see "pset7 Code Xiner Zhou.py" for details.

Short description of code as a whole: The main function "Greedy" calls 3 functions: genRandGraph(graph), BFS(graph, v), and sampleInfluence(G,S,m). The function genRandGraph(graph) takes input a graph and outputs a new random graph only has realized edges and nodes that have at least 1 edges remained in the sampled graph. The function BFS(graph, v) takes a graph and starting node v, and returns a dict indicating reachable or not for all nodes. The function sampleInfluence(G,S,m) takes a graph G, a subset of nodes S, and sample size m as inputs, and returns an empirial estimate of influence using sampling. Finally, the Greedy algorithm takes a graph G, number of initial adopters n, and sample size m; and returns the set of initial adopters S that maximize influence, as described in Q3.